Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Mallick and Yi, J Biomet Biostat 2013, 4:4
DOI: 10.4172/2155-6180.1000e124

Editorial Open Access

Hierarchical Models for Genetic Association Studies

Himel Mallick and Nengjun Yi*

Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA

Keywords: Bayesian hierarchical models; GLM; Genetic association studies; QTL mapping; GWAS

Challenges in Genetic Association Studies

It is well known that complex diseases are caused by a network of interacting factors and therefore statistically it makes more sense to consider multiple predictors (genetic variants and environmental covariates) simultaneously than to consider a single SNP (single-nucleotide polymorphism) [1,2]. Such joint analyses not only improve the power for detecting causal effects by explaining a substantial fraction of trait variation but also potentially lead to increased understanding about the genetics of the underlying traits or diseases [3]. The dramatic increase in genetic discoveries involving complex diseases has provided great opportunities for statisticians to contribute critical concepts and methods to the field [2,3]. However, analysis and interpretation of genetic association studies for jointly handling multiple genetic and environmental variables involves interesting statistical challenges [4-6]. First, with multiple SNPs and environmental factors, there are many possible main effects and interactions, most of which are likely to be null or at least negligible, leading to high-dimensional sparse models [7]. In addition, there are many more possible interactions than main effects, requiring different modeling and parameterization for main effects and interactions [2]. Second, genetic association studies usually genotype SNPs with strong linkage disequilibrium (LD), introducing highly correlated variables [7]. Third, SNP data often include genotypes with low frequencies that create predictors with near-zero variation [7]. Finally, separation, which arises when a predictor or a linear combination of predictors is completely aligned with the outcome, is a common phenomenon in case–control genetic association studies [7-9]. These complications result in challenges in terms of statistical modeling and computation and thus sophisticated techniques are required to handle them.

Why Hierarchical Models?

Recent years have seen an affluence of new methods for genetic and genomic research. As Allison et al. [10] states, "Many of these are wonderfully creative and useful". Among them, a potentially attractive approach, which has revolutionized modern genetic research, is Bayesian hierarchical modeling. Many papers have shown the practical and theoretical advantages of using Bayesian methods for genetic association studies [6]. Hierarchical models, which are more easily interpreted and handled in the Bayesian framework, are no exceptions. They provide a rational and quantitative way to incorporate biological information facilitating fitting of a large number of variables and a range of possible genetic models in a single analysis [6]. Thus, using appropriate prior information on the coefficients hierarchical models attempt to solve the aforementioned problems by providing stable, regularized estimates unlike non-hierarchical models, which generally cannot handle many variables simultaneously and often tend to overfit [2].

Key Concepts in Bayesian Hierarchical Models

Continuous shrinkage priors

For genetic models with a large number of potential genetic variants and environmental covariates, it is reasonable to assume that most of the variables have null or weak effects on the phenotype, whereas only a few have noticeable effects [11]. Bayesian hierarchical models incorporate this idea by setting up shrinkage prior distributions on the coefficients that give each effect a high probability of being near zero. The shrinkage prior distributions usually have very heavy tails, which enable strong shrinkage of small coefficients while minimally shrinking large coefficients. A variety of continuous shrinkage priors have been proposed in literature and many of them have been adopted to QTL mapping and genetic association analysis [7,12]. Two most commonly used continuous shrinkage priors are Student's t-distribution and the double exponential distribution. With these shrinkage priors, the posterior mode estimates of the coefficients are the ridge-penalized estimate [13] and the lasso-penalized estimate [14] respectively.

Scale mixtures of normals

Both the double exponential distribution and the Student's t-distribution can be presented as a two level hierarchical model [12,15]. The first level assumes that the coefficients follow independent normal distributions with mean zero and unknown variances whereas the second level assigns independent prior distributions on the variances which themselves depend on the hyperparameters. The two-level hierarchical formulation has several advantages. First, it allows easy and efficient computation facilitating MCMC and EM algorithms; conditional on the variances the coefficients can be easily estimated [2]. Secondly, it offers easy interpretation of the model; the coefficient-specific variances result in different shrinkage amounts for different coefficients [7]. Thirdly, it is flexible enough to encompass most popular penalized regression procedures and new hierarchical models can be defined by using new priors for the variances or further modeling the hyperparameters [16].

Algorithms for Fitting Hierarchical Models

Estimating posterior modes

A variety of methods for computing the posterior mode have been developed using the EM (expectation–maximization) algorithm which takes advantage of the two- level hierarchical formulation [17]. These algorithms have been adapted to multiple QTL mapping and genetic association analysis [11,18-20]. In a generalized linear model framework, given the variances, the prior distributions can be included as additional 'data points' in the normal approximation of the likelihood function. Therefore, the coefficients can be estimated using the

***Corresponding author:** Nengjun Yi, Department of Biostatistics, Ryals Public Health Building 317F, University of Alabama at Birmingham, Birmingham, AL 35294, USA, Tel: (205) 934-4924; Fax: (205) 975-2540; E-mail: nyi@uab.edu

Received June 02, 2013; **Accepted** June 04, 2013; **Published** June 08, 2013

Citation: Mallick H, Yi N (2013) Hierarchical Models for Genetic Association Studies. J Biomet Biostat 4: e124. doi:10.4172/2155-6180.1000e124

standard iterative weighted least squares (IWLS) algorithm for fitting classical generalized linear models [8,9]. This computational strategy takes advantage of the standard algorithm and leads to a stable, flexible and easily implementable computational tool [2,7]. The hierarchical generalized linear models with Student-t priors on the coefficients include various methods as special cases that have been designed to handle problems encountered in interacting QTL and association studies and therefore, can flexibly deal with various types of continuous and discrete phenotypes [2].

Sampling from continuous posterior distribution

On the other hand, taking advantage of the scale mixture representation of the shrinkage priors, various MCMC algorithms have been developed, many of which have recently been adapted to multiple QTL mapping and genetic association analysis [1,12,16,21-27]. Since all the priors for regression coefficients are conditionally Normal, a simple and unified scheme can be developed to update the coefficients regardless of the specific prior distributions on the variances [2]. For the Student-t and double exponential priors, the conditional posterior distributions of the variances have standard forms and thus can be easily sampled using MCMC [2,12]. Also, the hyperparameters can be assigned appropriate hyperpriors so that they can be updated along with other parameters or estimated based on empirical Bayes using marginal maximum likelihood [16]. These methods can simultaneously fit many correlated variables and can distinguish important effects from a large number of correlated variables [2,7].

Software Implementation

The above-mentioned algorithms for fitting Bayesian hierarchical models for genetic association studies can be carried out in a freely available R package BhGLM, which can be downloaded at http://www.ssg.uab.edu/bhglm/. The package provides a unified framework for setting up and fitting Bayesian hierarchical GLMs, for numerically and graphically displaying the results, with special emphasis on genetic association studies and QTL mapping. The functions in BhGLM are particularly useful for complicated genetic data analysis, e.g., QTL mapping in experimental crosses, genetic association studies for rare and common variants, prediction of complex diseases and traits, gene-set and pathway analysis, haplotype association analysis, gene-gene (GXG) and gene-environment (GXE) interactions, etc. Moreover, the methods can be used for general data analysis as well as for analyzing high-dimensional and correlated data arising from other disciplines. Some important functions in the package include but not limited to bglm (for fitting Bayesian hierarchical GLMs), bglm.ex (extensions of Bayesian hierarchical GLMs), bglm.selection (variable selection for Bayesian hierarchical GLMs), bpolr (Bayesian hierarchical ordered logistic or probit regressions for ordinal response), make.haplo (for creating a design matrix for haplotypes), make.inter (for making design matrix of interactions (GXG and GXE)), make. main (to make design matrix of main effects from genotypic data of genetic markers), plot. bglm (for graphically summarizing Bayesian hierarchical GLMs fits), predict.bglm (to make predictions for Bayesian hierarchical GLMs), summary.bglm (to summarize Bayesian hierarchical GLMs fits), etc. Other functions and additional details can be found at http://www.ssg.uab.edu/bhglm/.

Concluding Remarks

Although we have discussed continuous shrinkage priors due to their natural computational advantages, other shrinkage priors such as the spike and slab priors (discrete, two-component mixture distributions) [28] and Bayesian non-parametric hierarchical methods based on Dirichlet and other priors [29-32] are also common. Other continuous shrinkage priors include: the horseshoe prior [33]; the generalized double-Pareto model [34]; the orthant normal prior [22]; the mixture of uniform prior [35]; the Laplace-Poisson model [36] and the hypergeometric inverted-beta model [37]. The Bayesian hierarchical models discussed above can simultaneously analyse many covariates, main effects of numerous loci, epistatic and GXE interactions. These methods can return not only point estimates but also interval estimates of all the parameters, and offers a natural means of assessing model uncertainty [2]. For large-scale genetic data, a preliminary analysis is recommended to weed out unnecessary variables, or a variable selection procedure should be assessed to build a parsimonious model that only includes the most important predictors [21]. The only disadvantage of the Bayesian hierarchical approach is the intensive computation, which becomes more rigorous for high dimensional genetic data. Nevertheless, Bayesian hierarchical models attempt to overcome the associated challenges in high dimensional genetic data analysis by simultaneous variable selection and reliable coefficient estimation [16,26]. As noted by Zhou [3], statistical methods must evolve naturally to meet the challenges of sequence data and to resolve the missing dark matter of genetic epidemiology. Therefore, more coherent and sophisticated approaches based on hierarchical models that can combine external biological information and results across studies (meta analysis), across SNPs in genes and/or across gene pathways will further improve our understanding.

Acknowledgements

This work was supported in part by the research grants NIH 5R01GM069430-08 and U01 NS041588.

References

1. Li J, Das K, Fu G, Li R, Wu R (2011) The Bayesian lasso for genome-wide association studies. Bioinformatics 27: 516-523.
2. Yi N (2010) Statistical analysis of genetic interactions. Genet Res (Camb) 92: 443-459.
3. Zhou H, Sehl ME, Sinsheimer JS, Lange K (2010) Association screening of common and rare genetic variants by penalized regression. Bioinformatics 26: 2375-2382.
4. Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. Nat Rev Genet 5: 251-261.
5. Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet 86: 6-22.
6. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10: 681-690.
7. Yi N, Ma S (2012) Hierarchical Shrinkage Priors and Model Fitting for High-dimensional Generalized Linear Models. Stat Appl Genet Mol Biol 11.
8. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis. (2ndedn) Chapman & Hall/CRC, Florida, USA.
9. Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York, USA.
10. Allison DB, Visscher PM, Rosa GJM, Amos CI (2009). Editorial: Statistical genetics & statistical genomics: Where biology, epistemology, statistics, and computation collide. Comput Stat Data Anal 53: 1531-1534.
11. Yi N, Liu N, Zhi D, Li J (2011) Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. PLoS Genet 7: e1002382.
12. Yi N, Xu S (2008) Bayesian LASSO for quantitative trait loci mapping. Genetics 179: 1045-1055.

13. Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12: 55-67.

14. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58: 267-288.

15. Griffin JE, Brown PJ (2005) Alternative prior distributions for variable selection with very many more variables than observations. University of Warwick. Centre for Research in Statistical Methodology, Coventry, UK.

16. Kyung M, Gill J, Ghosh M, Casella G (2010) Penalized regression, standard errors, and Bayesian lassos. Bayesian Anal 5: 369-412.

17. Figueiredo MA (2003) Adaptive sparseness for supervised learning. IEEE Trans Pattern Anal Mach Intell 25: 1150-1159.

18. Xu S (2010) An expectation–maximization algorithm for the Lasso estimation of quantitative trait locus effects. Heredity 105: 483-494.

19. Yi N, Banerjee S (2009) Hierarchical generalized linear models for multiple quantitative trait locus mapping. Genetics 181: 1101-1113.

20. Yi N, Zhi D (2011) Bayesian analysis of rare variants in genetic association studies. Genetic epidemiology 35: 57-69.

21. Banerjee S, Yandell BS, Yi N (2008) Bayesian quantitative trait loci mapping for multiple traits. Genetics 179: 2275-2289.

22. Hans C (2011) Elastic net regression modeling with the orthant normal prior. J Am Stat Assoc 106: 1383-1393.

23. Li J, Zhang K, Yi N (2011) A Bayesian hierarchical model for detecting haplotype-haplotype and haplotype-environment interactions in genetic association studies. Hum Hered 71: 148-160.

24. Li Q, Lin N (2010) The Bayesian elastic net. Bayesian Analysis 5: 151-170.

25. Mutshinda CM, Sillanpaa MJ (2010) Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics 186: 1067-1075.

26. Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc 103: 681-686.

27. Sun W, Ibrahim JG, Zou F (2009) Variable selection by Bayesian adaptive lasso and iterative adaptive lasso, with application for genome-wide multiple loci mapping. The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series, Working Paper 10.

28. Yi N, Shriner D (2008) Advances in Bayesian multiple quantitative trait loci mapping in experimental crosses. Heredity 100: 240-252.

29. Dunson DB (2010) Nonparametric Bayes applications to biostatistics. Bayesian nonparametrics 28: 223.

30. Dunson DB (2009) Bayesian nonparametric hierarchical modeling. Biom J 51: 273-284.

31. Lijoi A, Prunster I (2010) Models beyond the Dirichlet process. Bayesian nonparametrics 28: 80-136.

32. Teh YW, Jordan MI (2010) Hierarchical Bayesian nonparametric models with applications. Bayesian Nonparametrics: Principles and Practice 28: 158-207.

33. Carvalho CM, Polson NG, Scott JG (2010) The horseshoe estimator for sparse signals. Biometrika, 97: 465-480.

34. Armagan A, Dunson DB, Lee J (2013) Generalized double Pareto shrinkage. Stat Sin 23: 119-143.

35. Knurr T, Laara E, Sillanpaa MJ (2011). Genetic analysis of complex traits via Bayesian variable selection: the utility of a mixture of uniform priors. Genetics Research 93: 303-318.

36. Chen X, Wang JZ, McKeown MJ (2011) A Bayesian Lasso via reversible-jump MCMC. Signal Processing 91: 1920-1932.

37. Polson NG, Scott JG (2012) Good, great, or lucky? Screening for firms with sustained superior performance using heavy-tailed priors. Ann Appl Stat 6: 161-185.